scientific reports

Check for updates

OPEN

Deep learning deciphers the related role of master regulators and G-quadruplexes in tissue specification

Artem Bashkatov[1], Andrey Andreasyan[1], Dmitry Konovalov[1], Alan Herbert[1,2]✉ & Maria Poptsova[1]✉

G-quadruplexes (GQs) are non-canonical DNA structures encoded by G-flipons with potential roles in gene regulation and chromatin structure. Here, we explore the role of G-flipons in tissue specification. We present a deep learning-based framework for the genome-wide G-flipon predictions across 14 human tissue types. The model was trained using high-confidence experimental maps of GQ-forming sequences and ATAC-seq peaks, conjoined with the location of RNA polymerase, histone marks, and transcription factor binding sites. The training dataset for the DeepGQ model was derived from EndoQuad level 4–6 GQs. Model predictions were subsequently validated against the comprehensive EndoQuad dataset (levels 1–6) to optimize the whole-genome prediction threshold. To identify tissue-specific regulatory patterns, we classified GQ promoter predictions as either 'core' or 'tissue-specific'. We identified a notable overlap between predicted unique tissue-specific GQ sites and master regulatory genes (MRGs), tissue-specific DNase-hypersensitivity sites, and proteins that modulate R-loop formation. Collectively, the findings highlight the transactions between MRG and G-flipons intermediated by RNA: DNA hybrids associated with tissue specification.

Keywords G-quadruplex, Flipons, R-loops, Tissue differentiation, Chromatin, Deep learning

The role of G-quadruplexes (GQs) in regulating genomic programs has been extensively studied. GQs in promoters are reported to participate in enhancer-promoter interactions, splicing regulation, chromatin assembly, DNA replication, recombination, and repair. GQs are encoded by sequences called flipons that cycle between different DNA conformations. These transitions occur under physiological conditions and facilitate the dynamic formation of transcriptionally active condensates[1].

Many whole-genome experiments have been used to detect GQ, including ChIP-seq[2], CUT&Tag[3] G4-seq[4], permanganate nuclease footprinting[5] and Kethoxal-assisted single-stranded DNA sequencing[6]. The results have been aggregated in the EndoQuad database[7] with confidence levels from 1 to 6 assigned according to the number of experiments that support GQ formation at a particular locus.

Computational approaches have also been developed to gain a deeper understanding of the dependence of GQ formation on DNA sequence and chromosomal location. These models use genome-wide information to elucidate the underlying biology and predict the context-specific outcomes. In the field of deep learning, several models have been evaluated. Initially, studies employed gradient boosting approaches[8] then later used convolutional/recurrent neural network (CNN/RNN)-based models next[9,10]. Now, large language models based on DNABERT are preferred[11]. All of these approaches can extend a very restricted set of experimental data to whole-genome predictions. These models were trained only on DNA sequence, as this data was readily available. Overall, transformer-based models have significantly improved model performance by incorporating sequence information that other architectures ignore. The results obtained underscore the importance of the surrounding region in determining GQ-forming sites. Such effects reflect the evolutionary selection of flipon features.

Only a small number of experimental studies have examined the regulatory role of GQs in different tissues, with GQ found more frequently in neural tissues than in other organs[12]. Currently, the processes involved in regulating tissue-specific GQ formation are not well understood. We can, however, use epigenetic codes to identify active GQ sites in any tissue, as it is likely that the same classes of machinery make these marks in every cell. We can also reduce the computational task by training the model on each tissue individually rather than

[1]International Laboratory of Bioinformatics, HSE University, Moscow, Russia. [2]InsideOutBio, Charlestown, MA, USA. ✉email: alan.herbert@insideoutbio.com; mpoptsova@hse.ru

all together. This approach is possible because many GQs are active in all tissues, with only a subset showing tissue-specificity. The common GQ helps the model identify those GQ that have a more restricted presence. We can then investigate those factors that predict tissue-specific formation of GQ. To tackle this problem, we have extended a deep learning approach first developed by us in DeepZ[13]. The DeepGQ model uses a combination of DNA sequence and omics features to predict GQ formation genome-wide. These predictions can also be applied to individual tissues using omic features measured in that tissue to identify tissue-specific sites of GQ formation.

To train DeepGQ to find epigenetic features associated with GQ formation, we used the set of high-confidence experimental GQs (levels 4–6) aggregated in the EndoQuad database. This set of G-flipons in EndoQuad was identified in a wide range of tissues with various methodologies. The training set used from EndoQuad is defined as those GQs detected in four or more independent datasets. DeepGQ then used ENCODE[14] epigenetic and transcription factor data to validate GQ predictions not only at the EndoQuad level 4–6 sites, but also those made genome-wide using other datasets. To optimize the performance of DeepGQ, the EndoQuad level 4–6 dataset was used to set the model parameters that maximized the true positive rate and minimized the number of false negatives. We used these thresholds to predict tissue-specific GQ based on ENCODE data from a tissue of interest. By comparing results for each tissue, we identified a universal "core" set of promoter GQs that are active in every tissue. We also found GQs that are active only in a subset of tissues, as well as others that are active in only a single tissue.

For each tissue, the whole-genome GQ predictions in promoters were then mapped to tissue-specific gene expression using orthogonal data from the GTEx[15] and TissueEnrich community resources. We subsequently tested whether promoter GQs active in a single tissue are associated with tissue-specific master regulator gene (MRG) binding sites. We then used DNase hypersensitivity sites (DHS) maps to test whether the MRG binding sites and tissue-specific GQ are colocalized within the same region of open chromatin.

To better understand factors involved in GQ formation, we focused on the mechanisms that create a region of single-stranded DNA (ssDNA) capable of folding into a G-quadruplex. One well-known method for forming ssDNA is by the displacement of one strand from a B-DNA duplex by an RNA. These structures are referred to as R-loops. RNA/DNA hybrids can be formed with both short and long RNAs, which can be produced either locally or at a different locus. We therefore used genome-wide data to investigate the overlap of G-flipons and R-loops. Overall, our results support a role for GQ and R-loop formation in MRG-dependent tissue-specific gene expression[16].

Results

Tissue-specific DeepGQ models based on tissue-specific omics features

Earlier, we developed the DeepZ approach that predicts functional Z-DNA regions based on sequence information from Z-DNA-specific ChIP-seq, chemical-footprinting experiments, and large sets of omics data[13]. The approach enables the extraction of meaningful information from limited and noisy datasets in a way that is not possible through the analysis of each experiment individually. In DeepZ, each omics feature was aggregated over all tissues to capture omics signal associated with Z-DNA formation. Here, we applied the same approach to predict G-quadruplexes, but with a focus on making tissue-specific predictions. The DeepGQ model is also trained for each tissue through a conjoint analysis of DNA sequences and omics features. Tissue-specific predications are then derived from omics features measured in a tissue of interest.

For the GQ training dataset, we chose EndoQuad database[7] as it is the most comprehensive collection of GQ detection datasets comprising more than 1200 G4-seq, G4 ChIP-seq and G4 CUT&Tag experiments. Each GQ in EndoQuad is assigned a confidence level of 1 to 6, corresponding to how many experiments confirmed GQ, where levels 1–3 correspond to one, two, and three experiments, and levels 4,5,6 to 4–5, 6–10, and >10 experiments. The entire EndoQuad dataset comprises approximately 390,000 GQs, of which around 133,000 are levels 4–6. As the training set, we chose EndoQuad levels 4–6, the most confident levels, provided that each GQ was detected in at least 4 experiments and also enriched for GQs common to many different tissues.

For each GQ and each tissue type, we assembled 14 tissue-specific omics datasets available from ChIP-Atlas, which include histone marks, transcription factor binding sites, ATAC-seq, and RNA Polymerase binding sites. The distribution of the type and number of omics features for each tissue is shown in Fig. 1B. The full list of omics features, including accession numbers for each experiment, is provided in Table S1.

In this computational approach, different types of architectures can be used, including CNN, RNN, hybrid CNN-RNN[13,17]and GNN[18]. The original implementation of DeepZ was based on RNN, but later we extended the approach to use GNN[18,19]. Here, we also utilized the RNN architecture, optimizing for the F1-score (Fig. 1A). We additionally employed a bi-directional LSTM module, as this method yielded the best performance/computational resources ratio among the tested deep learning models.

We trained 14 DeepGQ models, using 14 tissue-specific omics data (Fig. 1B,C). The train-test split was 75 – 25%. PR-curve on a test set (25% of EndoQuad, level 4–6) is given in Fig. 1D. To overcome the bias of EndoQuad for GQ shared by multiple tissues, we evaluated the performance of whole-genome DeepGQ maps on full EndoQuad, levels 1–6, and used this analysis to set a threshold for whole-genome predictions in each tissue examined. An example of how thresholds were set is shown for neural tissue is shown in Fig. 1E. Here, we generated whole-genome maps with threshold steps of 0.05 and calculated both the share of DeepGQ overlapping EndoQuad levels 1–6, and the share of EndoQuad overlapping DeepGQ (Fig. 1E). The intersection of the two plots corresponds to an optimum threshold of 0.25. The same DeepGQ overlap with EndoQuad levels 1–6 presented as light blue bars in Fig. 1E is depicted as PR-curve in Fig. 1D (also light blue curve).

Following training, the DeepGQ models allowed us to generate GQ annotations for 14 different tissues. Depending on tissue types, 70–80% of the DeepGQ predictions are found in EndoQuad, levels 1–6, with adipocytes and liver having more than 80% overlap (Fig. 1F). The reliability of the model was confirmed by

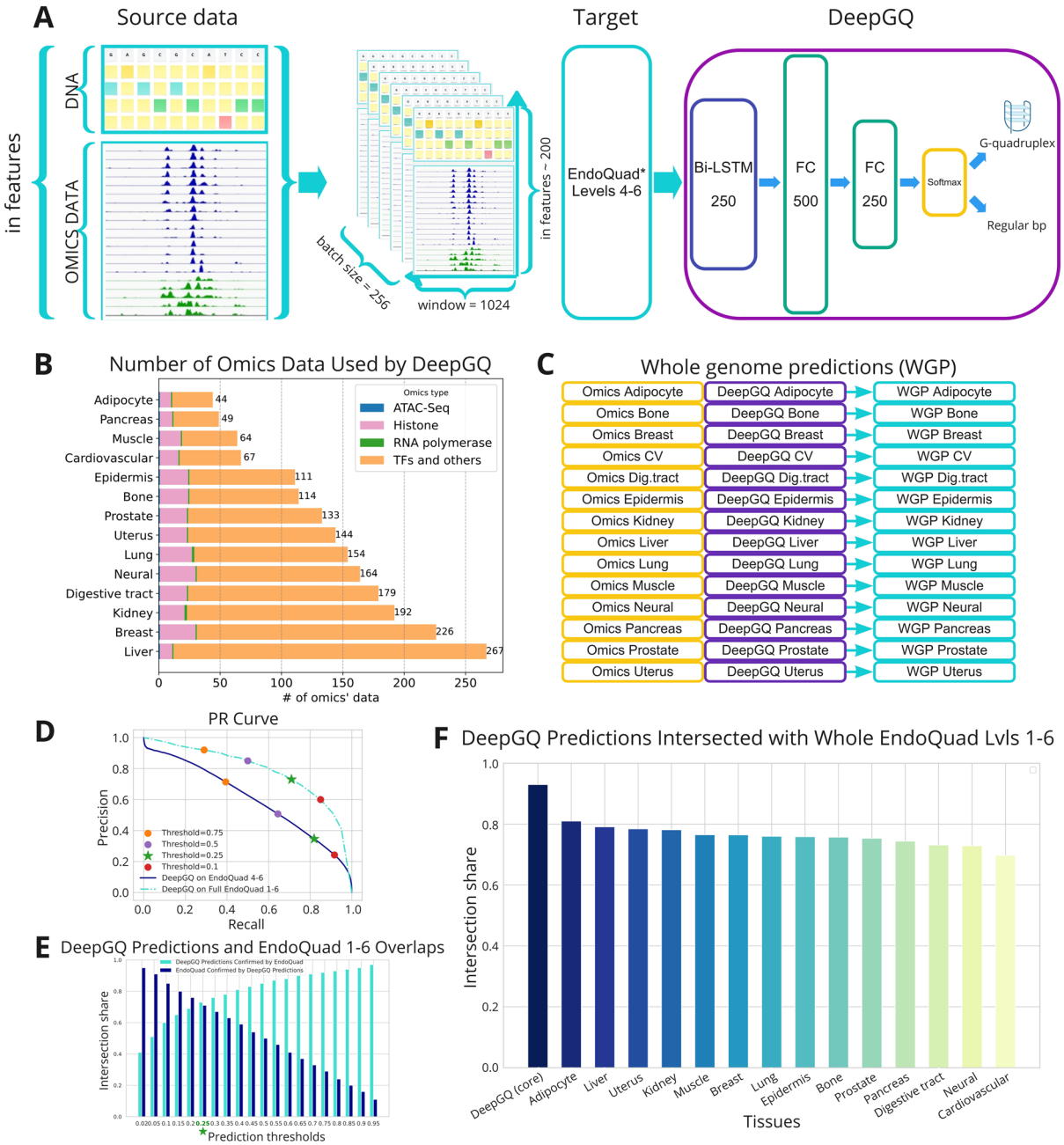


Fig. 1. DeepGQ model. (**A**) Schematic architecture of the DeepGQ model, which is based on RNN-LSTM, and as input, it takes a DNA sequence together with omics data from a tissue of interest. DeepGQ was trained on EndoQuad annotations, levels 4–6. (**B**) Omics features for different tissues. (**C**) Tissue-specific DeepGQ models. Separate DeepGQ models were trained using 14 tissue-specific omics features, with 14 GQ whole-genome predictions (WGP) generated. (**D**) The threshold used for predictions was set at 0.25 (marked with a star), as it gave the best balance between Precision/Recall to maximize the F1-score when evaluated using the entire set of EndoQuad annotations (levels 1–6). The PR curves and intersection ratios are shown for neural tissue at both EndoQuad levels 4–6 and levels 1–6. (**E**) The threshold of 0.25 yielded almost the same proportion of the DeepGQ predictions in EndoQuad as the proportion of EndoQuad annotations found in DeepGQ predictions. (**F**) Proportions of the DeepGQ tissue-specific predictions intersecting EndoQuad, levels 1–6.

benchmarking against other experimental datasets such as KEx[5] G4-seq[4], G4-ChIP[2] G4 CUT&Tag[3]and KAS-seq[6], highlighting its accuracy (Table S2). Performance metrics for all DeepGQ models are presented in Table S3.

By comparing GQ formation sites generated by each DeepGQ model, we annotated GQs that are common to many tissues, those present only in a subset of tissues, or unique to a particular tissue.

Identification of core and tissue-specific G-quadruplexes

To narrow our analysis, we focused on GQs located in the extended promoter regions as -2000 to +200 bp (hereinafter referred to as promoters) from the annotated transcription start sites (TSS). We then categorized the predicted GQs into 'core' (shared across all tissues) and tissue-specific subsets. Core GQs are derived from the overlap of all independent GQ predictions made for all tissues. The core GQ overlap boundaries represent the most upstream or downstream value obtained for the G-flipon predictions made for each promoter region (Fig. 2A). GQ not included in the core GQ set were labeled as "tissue-specific" (Fig. 2B). An example of the Venn diagram for 5 tissues (easy to visualize compared to 14) is given in Fig. 2B. Here we show a core set in the middle and tissue-specific subsets. As can be seen, many of the subsets are not unique for any particular tissue as they contain GQs belonging to two, three, four, etc., tissues. Only a small fraction of the predictions are unique to one of the 14 tissues. Nevertheless, as we show below, these subsets contain the same GQ signature as the core GQ.

We then used the GQ predictions unique to a specific tissue to generate a list of associated genes for further evaluation (see the complete list in Table S4). It is worth noting that 'tissue-specific' genes will also have core GQ predictions in their promoters. The presence of both GQ types in a promoter is not surprising, as core GQs likely play general roles in timing the onset of transcription and the resetting of promoters[20].

Functional enrichment of tissue-specific GQ-associated genes

Unique tissue-specific GQ-associated genes were evaluated using TissueEnrich[21]GTEx (Genotype-Tissue Expression)[15] and DAVID (Database for Annotation, Visualization, and Integrated Discovery) Gene Ontology tools[22] (Fig. 2C–E, Table S5). These analyses confirmed tissue-specific enrichment in functional pathways that are relevant to the tissues analyzed. For instance, genes with neural-specific GQ predictions in promoters showed significant enrichment in neuronal processes, supported by expression data from the GTEx portal. TissueEnrich-identified gene sets exhibited elevated expression in their corresponding tissues, as illustrated in Fig. S1. Interestingly, around twice as many neural-specific GQs were in the EndoQuad 1–3 set as in the 4–6 set. The result is consistent with their tissue-specific nature. The few experimental replications of this data explain their EndoQuad level 1–3 classification (Table S6), not that the input data into EndoQuad was unreliable.

Association of G-quadruplexes with master regulator genes

We investigated the relationship between unique tissue-specific GQ predictions and master regulator genes (MRGs). MRGs were selected based on their annotation in the literature as key regulators of tissue-specific differentiation[23]. We first identified MRG binding sites in the promoters of the tissue-specific genes annotated by the Human Protein Atlas resource (HPA)[24]. We then extracted the G-flipons predicted within this set of promoters and tested whether they were associated with MRG binding sites. We found significant GQ enrichment in promoters bound by MRG, as shown by the vertical red lines in Fig. 3A–C, when compared to the distribution generated by the random permutation of G-flipons with promoters from neurally expressed HPA genes. For example, 95% of NEUROD1 binding site-associated promoters in neural tissues contained at least one unique GQ, which is far higher than expected by chance (Fig. 3A). Similar enrichment patterns were observed for many, but not all other MRGs. For example, a significant association was found between MYOD1 in muscle and HNF1A in liver, with predicted unique GQs in tissue-specific promoters. Detailed results are presented in Fig. 3B–D and Table S7.

G-flipons present in a subset of tissues likely act by facilitating tissue formation by MRGs (Fig. 3E). If so, they play a supporting role in tissue specification. For example, this class of G-flipons may help modulate differentiation during the later stages of tissue development. In this situation, MRGs initiate a tissue-specific commitment that depends on the activation of a subset of the available GQs within the lineage by other transcription factors. The transcription factors involved in these later steps may also be present in a subset of tissues. An example is SOX2, a protein that, on its own, is not sufficient to initiate the tissue differentiation process. Instead, SOX2 facilitates the tissue-specific expression of genes by its ability to enable the assembly of complexes on different promoters in different tissues[25].

Formation of G-quadruplexes and R-loops

Formation of G-quadruplexes has been associated with R-loop formation[26]. Here, an R-loop arises when an RNA displaces the GQ-forming strand of DNA to disrupt the DNA helix. Previous studies have mapped the DNA/RNA hybrids formed by R-loops[16]. Here, we utilized a whole-genome map of R-loops, as detected by sequencing of DNA: RNA hybrid immunoprecipitates (DRIP-seq). The data is from a study of the APOBEC3B protein, an enzyme that performs cytosine-to-uridine editing of DNA. Editing by APOBEC3B is associated with the GQ-dependent C-to-G transversion in tumors, and was reported in this study to play an essential role in the resolution of R-loops[27]. We validated a significant overlap of the R-loop map with GQ annotated in EndoQuad: 68.33% for levels 4–6 and 19.54% for levels 1–3 (Table S6). Binding sites for APOBEC3B were also enriched in overlaps with both EndoQuad GQ (levels 4–6) and DeepGQ core predictions (58.81% and 57.96%, respectively). Other proteins are also known to bind GQ, such as the DHX36 helicase, which has a role in resolving both GQ and R-loops[28]. Both DeepGQ core predictions and EndoQuad 4–6 were enriched for binding of the helicase DHX36 (84.86% and 81.95%, respectively)[29] (Table S6). These findings confirmed colocalization of GQ with R-loops.

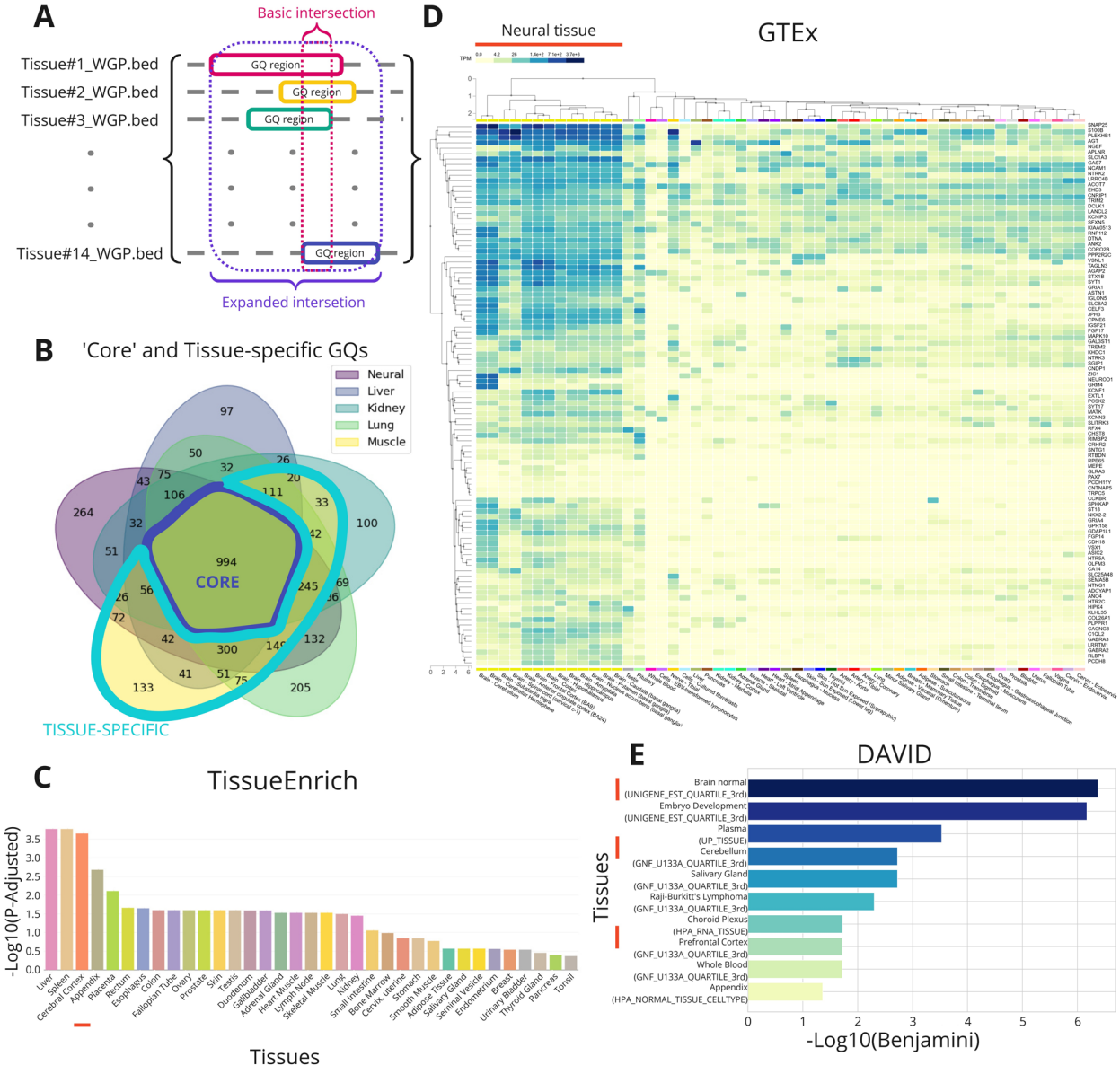
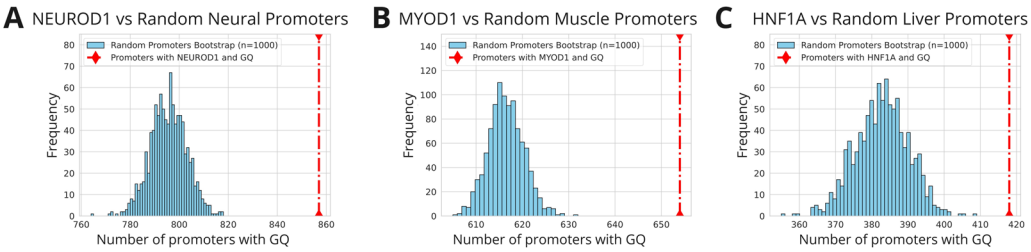


Fig. 2. Tissue-specific DeepGQ predictions. (**A**) Selection of "core" and tissue-specific quadruplexes. To distinguish the predictions that are common to all the tissues ('core') we intersected 14 whole-genome predictions (WGP). Intervals that overlap across all tissues simultaneously (Basic intersection) are extended to their farthest boundary (Extended intersection). Such Extended intersection intervals constitute the core dataset. Some GQ predictions are not "core" but common to a subset of tissues. Other GQ predictions are unique to the tissue analyzed. All genes with at least one unique tissue-specific GQ prediction in their promoter were included in each tissue-specific gene set. (**B**) An example of core and unique tissue-specific GQ predictions. Here we provide a Venn Diagram for 5 tissues. (**C**) Tissue-specific gene expression for promoters with a unique GQ prediction in neural tissues. The data was extracted with the TissueEnrich tool. The neural/brain-specific gene sets are underscored in red. (**D**) GTEx expression of genes highlighted by TissueEnrich as neural-specific. (**E**) Neural-specific gene sets were analyzed via the DAVID annotation tool.

For unique neural-specific GQ predictions, the overlap of R-loops from DNA: RNA hybrid track was 36.84%. Neural-specific DeepGQ revealed a 25.11% overlap with APOBEC3B binding sites and 36.84 with DHX36 binding sites. We validated these as significant overlaps by permutation testing (Table S6). Overall, the findings with a variety of orthogonal datasets support the usefulness of the DeepGQ predictions in associated GQ and R-loop formation.

Previously, we have shown that G-flipons are enriched for conserved miRNA binding sites and that the promoters involved are enriched for genes involved in neural development[30]. These small RNAs were proposed to act by forming DNA: RNA hybrids with promoter sequences, thereby directing the docking of AGO1 and



D

Tissue	MRG name	HPA genes (N)	with MRG	with MRG and GQ	GQ share	p_value
Neural	NEUROD1	2148	905	857	0.95	< 0.001
Neural	ASCL1	2148	1664	1563	0.94	< 0.001
Neural	PAX6	2148	75	73	0.97	< 0.001
Neural	SOX2	2148	156	142	0.91	0.078
Neural	NANOG	2148	418	374	0.89	0.115
Neural	POU5F1	2148	178	159	0.89	0.219
Liver	E2F1	949	254	245	0.96	< 0.001
Liver	HNF1A	949	525	418	0.80	< 0.001
Liver	GATA4	949	295	236	0.80	0.001
Liver	HNF4A	949	861	635	0.74	0.021
Liver	CEBPA	949	614	449	0.73	0.398
Liver	FOXA1	949	689	497	0.72	0.791
Breast	ESR1	122	102	68	0.67	< 0.001
Breast	STAT5A	122	37	28	0.76	0.001
Breast	CEBPB	122	47	32	0.68	0.027
Breast	FOXA1	122	93	57	0.61	0.036
Breast	GATA3	122	91	56	0.62	0.038

E

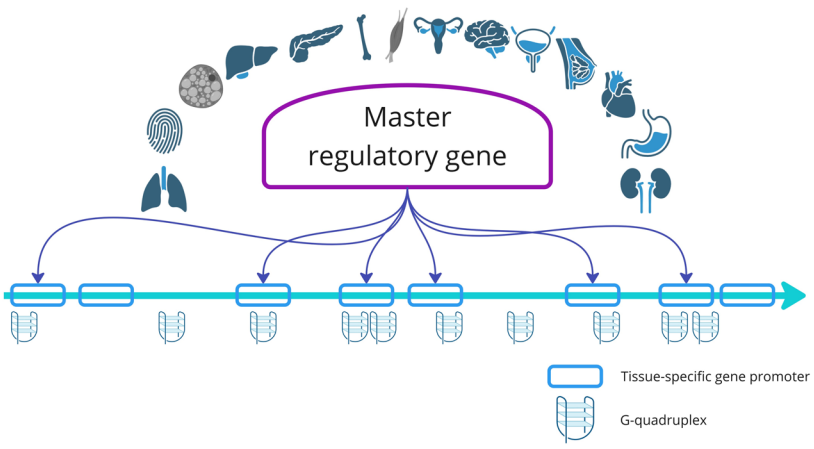


Fig. 3. Enrichment of master regulator binding sites with GQ. (**A**–**C**) Permutation tests for enrichment of master regulators binding sites with tissue-specific GQ promoters from the Human Protein Atlas (HPA). The red lines indicate the actual number of tissue-specific promoters that overlapped with master regulator binding sites and with GQ predictions simultaneously. The distribution corresponds to overlaps with 1,000 random selections of the same number of promoters (as the number of promoters with just MRG TFs) from the same tissue with GQ overlap. Permutation tests are presented for (**A**) neural master regulator NEUROD1, (**B**) muscle master regulator MYOD1, and (**C**) liver master regulator HNF1A. (**D**) Master regulators overlap with GQ tissue-specific genes, as indicated by corresponding p-values from permutation tests for three tissues: neural, liver, and breast. (**E**) Master regulator gene binding sites, located in the promoters of relevant tissue-specific genes, contain significantly more GQ predictions than random promoters.

AGO2 proteins to those promoters. We therefore examined the overlap between GQ, MRG binding sites, and those for AGO1 and AGO2 proteins. We found a significant overlap between the majority of the tracks (Fig. 4A). Most of the intersections, when assessed by permutation testing, showed a p-value < 0.001. Notably, we received significantly higher intersection rates when performing permutation tests only within promoter regions (-2000 bp to + 200 bp), compared to the whole genome. Datasets were not available to evaluate the role of long noncoding RNAs[31]or ADAR1[32], in R-loop dependent GQ formation.

We further investigated the overlap of both GQs and MRG binding sites with DHS to examine whether both features were colocalized in the same region of open DNA. We used the DHS initially mapped by the ENCODE project in 125 human cell and tissue types and later extended to 733 human biosamples, comprising 438 cell and tissue types and states[33,34]. The ENCODE DHS have a median length of 196 bp with cores having an average width of 55 bp (median 38 bp). Overall, 94.5% of ENCODE transcription factor binding sites fall within DHS. As we observed with GQ predictions, some DHS sites are tissue-specific, while others are common to several cell types[34]. Notably, our results show a remarkable overlap (Fig. 4B–E) of MRG and unique tissue-specific GQ within the same tissue-specific DHS. These findings provide additional biological validation of DeepGQ's predictive power and a role for G-flipons in modulating gene expression by MRGs.

Discussion

Our study provides a comprehensive framework for predicting tissue-specific G-quadruplex (GQ) structures across the human genome, utilizing an integrative deep learning approach enhanced with multi-omics data. By incorporating nucleotide-level one-hot encoded sequences and four major types of omics data (RNA polymerase, histones, ATAC-seq, and transcription factors), we developed a robust predictive model capable of delineating tissue-specific regulatory mechanisms. We provide a high-resolution genome-wide GQ mapping where the DeepGQ model generated a genome-wide GQ map, predicting GQ presence at single-nucleotide resolution across 14 human tissues. Validation using external datasets, including G4-seq, G4-ChIP, and CUT&Tag, confirmed the high accuracy of these predictions. This approach significantly advances our understanding of the role of GQs in gene regulation.

The DeepGQ approach enabled the generation of tissue-specific regulatory insights. By categorizing GQ predictions into 'core' and 'tissue-specific' sets, we identified distinct GQ regulatory patterns in gene promoters. Genes with promoters containing unique, tissue-specific GQs were shown to align with known tissue-specific functions, as demonstrated through DAVID and TissueEnrich analyses. Further analysis using GTEx data corroborated the role of neural-specific GQ-enriched genes in driving elevated expression levels in neural tissues. The results were replicated in other tissues.

Our findings also underscore a role of R-loop formation in regulating G-flipon conformation. Both AGO protein complexes and editing by APOBEC3B and ADAR enzymes, as well as helicases such as DHX36 and DHX9, have been previously shown to modulate R-loop formation. The docking of all these enzymes to GQ helps localize their actions. The association of GQ with R-loop formation is well known and is expected to increase with transcription. The consensus is that R-loops, whether formed by small or long RNAs, drive GQ formation rather than the other way around[16]. In this scenario, the formation of an ssDNA capable of folding into a GQ is powered by RNA transcription. Other data suggest that many promoters are bidirectional, with both sense and antisense transcripts, indicating that R-loop formation can be promoted by both coding and noncoding transcription[30,35,36].

Previous studies have also implicated the transcription of small enhancer RNAs (eRNAs) in GQ formation[37]. During development, the G-flipons modulated by RNAs may originate from endogenous retroviruses (ERVs) located upstream of the promoters used following tissue differentiation. Here, the formation of R-loops depends on the ERV promoter. Another possibility involves the sequence-specific binding of small RNAs to modulate the formation of GQ, including those arising from loci elsewhere in the genome. Indeed, we have previously demonstrated a highly significant overlap between G-flipon sequences and binding sites for conserved miRNA seed sequences in the promoters of developmental genes, particularly those involved in neural processes.

Our findings highlight significant enrichment of unique tissue-specific GQs in the promoter regions of genes bound by tissue-specific MRGs. For example, NEUROD1 in neural tissues exhibited a strong association with GQ structures in its binding sites, with a higher-than-random enrichment validated by permutation tests. Similar patterns were observed for other MRG-tissue pairs, indicating that GQs are integral to tissue-specific regulatory networks. Most notably, we observed an overlap between tissue-specific GQ mappings and tissue-specific DHSs, suggesting that these features are closely related.

How tissue-specific G-flipons and MRG interact remains an open question. In general, G-flipons can facilitate tissue specification by MRGs in multiple ways. By forming GQ, G-flipons can maintain regions of open chromatin, thereby preventing the occlusion of TF docking sites by chromatin. GQs also facilitate the formation of MRG-dependent enhancer-promoter condensates and enable the resetting of promoters after transcriptional bursts. Such processes allow cells to transition through various chromatin states in response to ongoing perturbations[12,38]. GQ can also localize sequence-specific TFs to promoters[39]. These processes can arise from R-loop formation or initiate R-loop formation through transcription or by unmasking RNA-binding sites.

The exact timing of R-loop formation, GQ formation, and MRG docking during development likely depends on context. Different possible scenarios exist. DNA: RNA hybrids formed by small RNAs could initiate GQ formation and open chromatin formation, allowing MRG to bind. Another possibility is that GQ formation, due to retroelement transcription, creates regions of open chromatin in the surrounding regions, thereby exposing MRG binding sites. Alternatively, pioneering MRG could displace nucleosomes to create open chromatin, initiate transcription, and promote R-loop and GQ formation. All these processes potentially enhance tissue-specific gene expression.

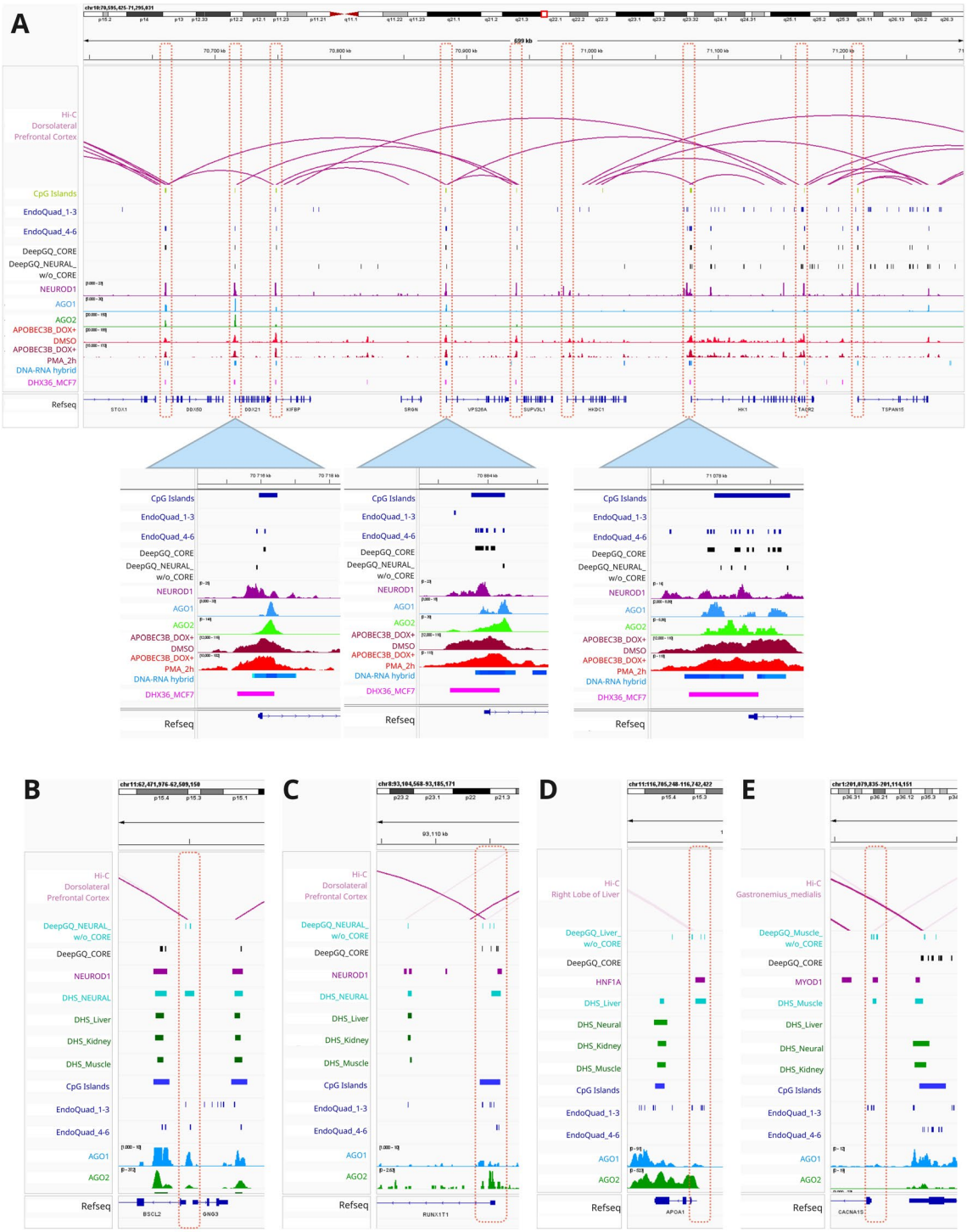


Fig. 4. Formation of G-quadruplexes. (**A**) Three examples of genes (DDX21, VPS26A, HK1), illustrating all of the tracks' data within their promoters for neural tissue, including: Hi-C Prefrontal cortex track, CpG islands, EndoQuad 1–3 and 4–6 levels, DeepGQ Core and DeepGQ Neural predictions, NEUROD1 (Neural MRG), AGO1, AGO2 binding sites, APOBEC3B DOX + DMSO and PMA tracks, DNA-RNA hybrid formation and DHX36 track. (**B**–**E**) Examples of four tissue-specific DHS maps for neural (**B**, **C**), liver (**D**), and muscle (**E**) tissues confirming overlap between tissue-specific GQ predictions, relevant tissue-specific MRG binding sites, and chromatin accessibility at sites of tissue-specific DHS.

Summary

The DeepGQ framework, therefore, provides a robust platform for genome-wide, tissue-specific GQ prediction and functional analysis. By integrating deep learning with omics data, it is possible to uncover significant associations between GQs, tissue-specific genes, and master regulators of tissue differentiation, allowing novel insights into the modulatory role of GQ in tissue-specific gene expression. We provide several metrics to validate the platform, highlighting a significant and important biological role for G-flipons.

Our integrative approach to predicting GQ structures, combining deep learning with omics data, represents a significant methodological advancement over earlier platforms. The refinement of prediction intervals and the incorporation of multiple data layers allowed for precise and biologically relevant predictions. The use of thresholds optimized for the F1 score ensured robust performance across diverse datasets.

These analyses underscore the critical role of GQ structures in gene regulation, particularly in tissue-specific contexts. The enrichment of GQs in promoter regions bound by MRGs suggests that GQs may facilitate specific regulatory interactions, enhancing transcriptional precision. The identification of core GQ regions, conserved across tissues, further highlights their potential involvement in fundamental cellular processes. The DeepGQ predictions enable identification of G-flipon sites for further experimental interrogation of how RNAs modulate GQ formation to regulate gene expression.

Limitations and future directions

While our study provides valuable insights, certain limitations remain. The reliance on publicly available omics datasets introduces variability due to differing experimental protocols. Additionally, the use of the hg19 reference genome may overlook variations in GQ structures across populations or updated genome builds. Future research could address these limitations by incorporating single-cell data and utilizing more recent genome assemblies.

Our study demonstrates a strong association between GQs and tissue-specific gene regulation, providing specific cases where wet laboratory experiments can confirm causality. Techniques such as CRISPR-mediated GQ manipulation or G4-specific ligands could offer deeper insights into the functional roles of GQs. Basing these studies on DeepGQ predictions provides a strategy for the efficient utilization of available benchside resources.

Overall, the DeepGQ framework offers a powerful tool for exploring tissue-specific GQ distributions and their regulatory implications. By elucidating the interplay between GQs, gene promoters, and MRGs, this study advances our understanding of genomic regulation. Our approach opens new avenues for investigating the functional roles of GQs in health and disease. Future applications of this framework may include its adaptation for use with other organisms or its integration into personalized medicine strategies for targeting tissue-specific regulatory elements.

Methods

DeepGQ model for tissue-specific quadruplex prediction

Similar to other deep learning approaches, including DeepZ[13]we developed the DeepGQ model to predict tissue-specific G-quadruplexes (GQs) using a deep learning framework. The model utilizes the hg19 genome encoded as one-hot DNA sequences, supplemented with omics data from next-generation sequencing (NGS) experiments. The dataset includes four major omics data types: RNA polymerase, histone modifications, ATAC-seq, and TFs. Tissue-specific omics data for 14 human tissues were sourced from the ChIP-Atlas[40] database as BED interval files. The analysis focused on nucleotide-level predictions, and the EndoQuad G-quadruplex dataset[7] (confidence levels 4–6) served as the ground truth.

To process the data, we merged omics intervals from multiple experiments into consolidated BED files and combined them with the one-hot encoded DNA sequences. The resulting genome-wide Boolean arrays represent the presence (1) or absence (0) of GQs. The dataset was segmented into windows and batches, as illustrated in Fig. 1A. Uninformative datasets (< 10 kB) were excluded to optimize computational efficiency. The availability of omics data across tissues is summarized in Fig. 1B and Table S1.

Model architecture and evaluation

We tested several deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks with long short-term memory (RNN-LSTM) units. Consistent with results from the DeepZ model, the RNN-LSTM architecture outperformed alternatives and was selected for further analysis. Detailed architectural parameters and performance metrics are provided in Fig. 1A, Supplementary Methods, and Table S3.

To determine the optimal classification threshold, we prioritized maximizing the F1-score. The evaluation was performed on the EndoQuad dataset (confidence levels 4–6) used for training, as well as the full dataset (levels 1–6) for independent validation, the model had not been exposed to during training. A threshold of 0.25 yielded the best performance for GQ predictions, when assessing on the full EndoQuad 1–6 (Fig. 1D,E). Post-prediction, raw genome-wide predictions were refined by eliminating intervals shorter than 10 base pairs and merging adjacent intervals separated by < 20 base pairs. The final output consisted of 14 BED files containing tissue-specific predictions, available in Supplementary Material. Model metrics are enclosed as Table S3.

Predictions were validated against a hg19 gene promoter map derived from the UCSC database[41]. We chose extended promoter regions of -2000 to + 200 bp from the transcription start site (TSS) to include proximal enhancer that are often located at 1-2kB upstream. The overlap between GQ predictions and promoter regions was evaluated using Bedtools.

Core and tissue-specific predictions

To distinguish between core and tissue-specific GQs, genome-wide predictions were intersected across all 14 tissues using Bedtools. Predictions common to all tissues formed the "core" set, which was expanded to

include the farthest overlapping intervals (Fig. 2A). Once formed, the 'core' and 14 tissue-specific GQ maps were evaluated against other available datasets and models (Table S2). As shown in the table, DeepGQ results align well with the complete EndoQuad dataset (all levels), which was not included in the training process (Fig. 1F). G4-seq, G4-ChIP, and CUT&Tag data also show strong overlap with the predicted GQ maps.

Genes with at least one core GQ in their promoter constituted the 'core' gene set, while others with no core and at least one tissue-specific GQ formed 'tissue-specific' gene lists (Fig. 2B). Table S4 contains comprehensive lists of core and tissue-specific genes.

Functional enrichment analysis

The core and tissue-specific gene lists were evaluated using DAVID and TissueEnrich tools. Tissue-specific gene enrichment was confirmed using the TissueEnrich tool, with examples for neural tissue shown in Fig. 2C,E. TissueEnrich gene lists were further analyzed for expression profiles using GTEx portal data, which demonstrated elevated tissue-specific expression levels (Fig. 2D). The full TissueEnrich and GTEx results are available in Fig. S1 and Table S5.

Master regulator gene analysis

We further explored the association between G-quadruplexes (GQs) and tissue-specific regulatory mechanisms by analyzing master regulator genes (MRGs). First, we examined the interactions between MRGs and the relevant tissue-specific genes. In this analysis, we included not only genes with predicted GQ structures in their promoters but also those classified as tissue-enriched, group-enriched, or tissue-enhanced based on data from the Human Protein Atlas (HPA)[24] and UCSC[41] databases. For this analysis we extended promoter-proximal enhancer regions by 2 kB flanks to account for more regulatory landscape and searched for overlap in a region of −4 kb to +2 kb relative to the transcription start site (TSS).

We selected the neural master regulator gene NEUROD1 as an example to investigate the intersection of master regulator gene (MRG) binding site maps with neural gene promoter maps derived from the Human Protein Atlas (HPA). The HPA neural/brain gene promoter map included 2,148 unique genes. By intersecting this dataset with the NEUROD1 binding site map, we identified 905 promoters containing at least one NEUROD1 binding site. We then isolated all 905 neural gene promoters from the HPA and intersected this set with a neural-specific GQ prediction map. This analysis revealed that 857 of the 905 neural gene promoters intersecting with NEUROD1 binding sites (~95%) contained at least one GQ, a result significantly higher than random chance (Fig. 3A). To validate these findings, we performed a permutation test. Specifically, we selected an equivalent number of random promoters (905) from the neural HPA dataset and performed GQ intersection 1,000 times. Our analysis demonstrated that neural gene promoters with NEUROD1 binding sites are significantly enriched in GQs when compared to random hits.

The same approach was applied to other tissues and MRGs, including muscle (MRG MYOD1) and liver (MRG HNF1A), with results visualized in Fig. 3B,C. A broader analysis across 39 tissue-MRG experiments demonstrated that 67% exhibited significant GQ enrichment in MRG promoter regions compared to random controls. Detailed results for three tissues (neural, liver, and breast) are shown in Fig. 3D, with additional findings available in Table S6.

Tissue enrichment analysis

The enrichment of tissue-specific genes was done with the TissueEnrich application[21] limited to genes and tissues from the Human Protein Atlas. The data used for the tissue expression analyses described in this manuscript were obtained from the GTEx Portal on 10/15/2024. The tissue-specificity assessment was performed using the DAVID tool[42]. Tissue-specific gene lists were obtained from the[24] (proteinatlas.org) on 08/11/2024.

Data sources

The human genome hg19 sequence was downloaded from UCSC Genome Browser (https://genome.ucsc.edu/). Omics datasets used for building DeepGQ models were downloaded from ChIP-Atlas portal (https://chip-atlas.org/)[43] on 09/12/2024. The complete list of omics data for each tissue type is provided in Table S1.

The EndoQuad database was downloaded from the website (https://EndoQuad.chenzxlab.cn/)[7]. APOBEC3B tracks were obtained from GEO repository under GSE148581[27]). DHX36 track was obtained from GEO repository under GSM8333909[44]. Hi-C maps, AGO1-2 binding site tracks, DHS mapping, MRG binding sites, DNA-RNA hybrid track were downloaded from ChIP-Atlas portal (https://chip-atlas.org/)[43] on 02/12/2025.

Data availability

The code is freely available at https://github.com/hse-bioinflab and https://github.com/maleficar259/DeepGQ.

Received: 5 March 2025; Accepted: 16 June 2025
Published online: 02 July 2025

References

1. Herbert, A. *FLIPONS: the Discovery of Z-dna and Soft-wired Genomes* (CRC, 2024).
2. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin Immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **13**, 551–564 (2018).
3. Lyu, J., Shao, R., Yung, K., Elsasser, S. J. & P.Y. & Genome-wide mapping of G-quadruplex structures with cut&tag. *Nucleic Acids Res.* **50**, e13 (2022).

4. Chambers, V. S. et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).
5. Kouzine, F. et al. Permanganate/S1 nuclease footprinting reveals Non-B DNA structures with regulatory potential across a mammalian genome. *Cell. Syst.* **4**, 344–356e7 (2017).
6. Lyu, R. et al. KAS-seq: genome-wide sequencing of single-stranded DNA by N(3)-kethoxal-assisted labeling. *Nat. Protoc.* **17**, 402–420 (2022).
7. Qian, S. H. et al. EndoQuad: a comprehensive genome-wide experimentally validated endogenous G-quadruplex database. *Nucleic Acids Res.* **52**, D72–D80 (2024).
8. Sahakyan, A. B. et al. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).
9. Klimentova, E., Polacek, J., Simecek, P. & Alexiou, P. P. E. N. G. U. I. N. N. Precise exploration of nuclear G-Quadruplexes using interpretable neural networks. *Front. Genet.* **11**, 568546 (2020).
10. Barshai, M., Aubert, A. & Orenstein, Y. G4detector: convolutional neural network to predict DNA G-quadruplexes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (2021).
11. Poptsova, M., Herbert, A., Konovalov, D. & Umerenkov, D. Analysis of live cell data with G-DNABERT supports a role for G-quadruplexes in chromatin looping. *bioRxiv* (2024).
12. Hansel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).
13. Beknazarov, N., Jin, S. & Poptsova, M. Deep learning approach for predicting functional Z-DNA regions using omics data. *Sci. Rep.* **10**, 19134 (2020).
14. Encode project. – resource with results of huge number of bioinformatics experiments.
15. Consortium, G. T. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
16. Wulfridge, P. & Sarma, K. Intertwining roles of R-loops and G-quadruplexes in DNA repair, transcription and genome organization. *Nat. Cell. Biol.* **26**, 1025–1036 (2024).
17. Beknazarov, N., Konovalov, D., Herbert, A. & Poptsova, M. Z-DNA formation in promoters conserved between human and mouse are associated with increased transcription reinitiation rates. *Sci. Rep.* **14**, 17786 (2024).
18. Voytetskiy, A., Herbert, A. & Poptsova, M. Graph neural networks for Z-DNA prediction in genomes. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 3173–3178 (Las Vegas, 2022).
19. Alaeva, A. et al. OmiXAI: An ensemble XAI pipeline for interpretable deep learning in omics data. *bioRxiv* (2025).
20. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).
21. Jain, A., Tuteja, G. & TissueEnrich Tissue-specific gene enrichment analysis. *Bioinformatics* **35**, 1966–1967 (2019).
22. Dennis, G. Jr. et al. Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, P3 (2003).
23. Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nat. Rev. Mol. Cell. Biol.* **23**, 449–464 (2022).
24. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
25. Julian, L. M., McDonald, A. C. & Stanford, W. L. Direct reprogramming with SOX factors: masters of cell fate. *Curr. Opin. Genet. Dev.* **46**, 24–36 (2017).
26. Lee, C. Y. et al. R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. *Nat. Commun.* **11**, 3392 (2020).
27. McCann, J. L. et al. APOBEC3B regulates R-loops and promotes transcription-associated mutagenesis in cancer. *Nat. Genet.* **55**, 1721–1734 (2023).
28. You, H., Lattmann, S., Rhodes, D. & Yan, J. RHAU helicase stabilizes G4 in its nucleotide-free state and destabilizes G4 upon ATP hydrolysis. *Nucleic Acids Res.* **45**, 206–214 (2017).
29. Liu, T. et al. Genome-wide mapping of native co-localized G4s and R-loops in living cells. *Elife* **13** (2024).
30. Herbert, A. Flipons and small RNAs accentuate the asymmetries of pervasive transcription by the reset and sequence-specific microcoding of promoter conformation. *J. Biol. Chem.* **299**, 105140 (2023).
31. Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell. Biol.* **22**, 96–118 (2021).
32. Shiromoto, Y., Sakurai, M., Minakuchi, M., Ariyoshi, K. & Nishikura, K. ADAR1 RNA editing enzyme regulates R-loop formation and genome stability at telomeres in cancer cells. *Nat. Commun.* **12**, 1654 (2021).
33. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
34. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
35. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell.* **58**, 1101–1112 (2015).
36. Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome Res.* **31**, 2170–2184 (2021).
37. Harrison, L. J. & Bose, D. Enhancer RNAs step forward: new insights into enhancer function. *Development* **149** (2022).
38. Herbert, A. Flipons and small RNAs accentuate the asymmetries of pervasive transcription by the reset and sequence-specific microcoding of promoter conformation. *J. Biol. Chem.* **299** (2023).
39. Spiegel, J. et al. G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.* **22**, 117 (2021).
40. Zou, Z., Ohta, T., Oki, S. & ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Res.* **52**, W45–W53 (2024).
41. Perez, G. et al. The UCSC genome browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2025).
42. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
43. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **50**, W175–W182 (2022).
44. Lu, Z. et al. Rapid degradation of DHX36 revealing its transcriptional role by interacting with G-quadruplex. *Aggregate* **6**, e647 (2025).

Acknowledgements

This work was supported by the Ministry of Economic Development of the Russian Federation (code 25-139-66879-1-0003).

Author contributions

M.P. and A.H. conceived the study. A.B. developed DeepG models and performed the data analysis. A.A. fine-tuned for the best DeepGQ architecture. A.B., M.P. and A.H. wrote the manuscript. D.K. performed DeepGQ comparisons with other GQ maps. All authors analysed and discussed the results. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-07579-1.

Correspondence and requests for materials should be addressed to A.H. or M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.